Big Data and Digital Humanities

Jochen Tiepmar

Abstract In academic discourse, the term Big Data is often used incorrectly or not considered in relevant use cases. This paper investigates the term Big Data in the context of text oriented digital humanities and in the process shows that it is not necessarily an issue of big data sets. The goal is to provide a starting point or a guideline for researchers in the humanities to relate their work to the concept of Big Data. It may even show the reader that they might be working on a task that can be considered as Big Data even though the data set itself is comparatively small. As such, this paper should not be seen as a concrete solution to specific problems but as a general overview that is based on several years of practical research experience. This paper also argues that interoperability is one of the most prominent Big Data issues in text oriented digital humanities.

Jochen Tiepmar
Leipzig University, Institute for Computer Science,
✉ jtiepmar@informatik.uni-leipzig.de

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 5, No. 1, 2018

DOI 10.5445/KSP/1000087327/01
ISSN 2363-9881

1 Introduction

Defining the term Big Data is not trivial. The most obvious defining factor is the size of a data set, but this property can not be applied universally and depends on the domain context as well as data type specific properties or measurements. For instance, text volume can be measured as *number of tokens/documents* or *byte*. While a token or document count can often result in impressive and seemingly large numbers, the corresponding bytes are often not in an area that can be considered as large. Yet certain text mining analyses – like citation analysis – and use case specific circumstances – like a real time requirement – may result in workflows that are already too calculation expensive for technically small data sets. IBM suggests the 4 *V*s, data specific properties to help describe the Big Data relevance of a problem. These *V*s are *Volume*, *Veracity*, *Velocity* and *Variety*.

1.1 Volume

Volume is the most obvious aspect of Big Data and describes the size of a data set. The bigger a data set is, the more effort is required to process, share or store it. Especially medical applications like analysis of MRI images and simulations like weather models or particle systems can create and require large amounts of data. The increasing amount of digital and sometimes publicly available sensory information that is collected – for a vast number of examples, see works about *Smart Cities* or *Internet of Things* – will probably increase the need for solutions for size-related problems.

Usually, a data set is not characterized as a Big Data problem if smaller than at least 1 Terabyte, and since current standard database systems and hard drives are able to store and manage several terabytes of data without any major issues, most Big Data Volume problems deal with memory and not disk space. Information that is stored in memory can be accessed faster than that in disk drives, but it is lost when the system is shut down. Therefore, disk space is usually used to store, manage, and archive data sets while memory is usually used for more dynamic, analytical tasks. Memory is currently also more expensive – and, therefore, more limited – than disk space, which means that the memory requirements that

qualify as a Big Data problem are usually lower than disk-space requirements. An arbitrarily chosen estimated border value could be 100 Gigabytes.

In the context of text-oriented digital humanities, *volume* can also be used to refer to more information-related aspects like the number of tokens, sentences, or documents, as it is usually done for text corpora. Information-related size statistics can quickly result in seemingly big and impressive numbers while the required disk space stays relatively small. In the context of this analysis, Volume with a capitalized letter *V* refers to disk or memory space.

Table 1 illustrates this relationship for some of the biggest data sets (Deutsches Textarchiv (DTA), Geyken et al (2011); Textgrid, Neuroth et al (2011)) that were collected in the context of this work. The disk space is calculated based on the uncompressed data set that is available for download and usually includes additional markup, which implies that the actual text data Volume is usually smaller. The number of documents and tokens is calculated based on the data set. The document number is the number of individual files, and the tokens were delimited by the characters =*"*<.>*()[]{},:;*, *tab*, *newline*, and *whitespace*. Textgrid provides multiple documents as part of one XML file, namely the <TEICorpus> with several TEI documents. These documents were separated into individual files. The token and document count can differ from the official project statistics, because they include the XML markup. This is intentional, since the point is to illustrate the relation between the number of words in a set of files and their hard disk space and, for this comparison, it is more correct to include the markup as tokens as it also influences the file sizes.

Table 1: Text corpus statistics vs. hard disk space.

Text Corpus	Documents	Tokens	Disk Space
DTA	2,435	211,185,949	1.3 GB
Textgrid	91,149	232,567,480	1.8 GB
PBC[1]	831	289,651,896	1.9 GB

As Table 1 shows, the required disk space for text data is quite small even for comparatively big data sets. Problematic file sizes can usually only occur for

[1] Parallel Bible Corpus (PBC), Mayer and Cysouw (2014)

text data sets that include optical scans of the document pages, which shall not be considered as text data but as image data. The English Wikipedia can be considered as one of the largest online text collections. Yet, according to its own statistics,[2] as of February 2013, the size of the XML file containing only the current pages, no user or talk pages, was 42,987,293,445 bytes uncompressed (43 GB). It can be stated that storing and managing text data is not a Volume problem with respect to disk size. The data size is also not problematic with respect to setups that are designed to work in memory. At the time of writing, the current prices for 64 GB RAM based on Amazon.com range from 541.95 €[3] to 1,071.00 €,[4] which might be too expensive to consider this as standard hardware, but this is probably far from problematic for a project that is designed with the requirement of managing a Wikipedia-size text collection in memory.

It must be emphasized that this is not a phenomenon that occurs because the amount of data is still small, and, therefore, can be expected to change in the near future. Instead, it can be considered as a constant characteristic of the practical use of text data. Data sets in this context correspond to individual document collections that tend to include documents that share a certain set of properties like a specific author, language, time period, or any kind of thematic relation. *Das Deutsche Textarchiv* only includes German literature covering a relatively limited time frame, and the *Parallel Bible Corpus* only includes Bible translations. Even if a data set includes a wide array of parameter configurations it can always be distinguished from other data sets by its specific properties. It is highly unlikely that the trend for this kind of data is headed toward centralization. This characteristic is especially important in text analysis because, in order to research specific effects, it is important to eliminate the impact of unrelated variables. A token frequency trend analysis usually requires a monolingual text corpus to avoid effects like the German feminine noun article *die* being counted as the English verb *to die*. Even in more inclusive use cases like a global digital library, it can be counter-productive not to limit the content to books and include – for instance – Twitter data or collected forum discussions. Therefore, it can be stated that the relatively small disk or memory size required to manage only the text data is and will not be a Big Data-related problem because of

[2] https://en.wikipedia.org/wiki/Wikipdia:Size_of_Wikipedia#Size_of_the_English_Wikipedia_database

[3] HyperX FURY DDR4 HX421C14FBK4/64 RAM Kit 64GB (4x16GB) 2133MHz DDR4 CL14 DIMM

[4] Kingston KVR18R13D4K4/64 RAM 64GB (DDR3 ECC Reg CL13 DIMM Kit, 240-pin)

the purpose and characteristics of this kind of data. It is unlikely that the size of document collections is an issue that cannot be solved using present-day standard hardware.

If text content is considered **primary data**, then external annotations and annotated text content can be considered **secondary data**. Annotated text content can include information about Part-of-Speech tags, named entities, editorial notes, and much more. External annotations can include citation links or linked resources like audio or image snippets to specific text passages. Secondary data does not have to be associated with the original text and can also occur as word frequency tables, topic models, co-occurrence & collocation data, or in the form of any other analytical result format.

Secondary data in the context of text is usually the result of automated analytical processes or manual editing. Especially, the amount of information that is added by automated analytical processes can significantly increase the Volume of a data set. The amount of this kind of data depends on the analytical processes that are done and the results that are produced. A representative overview of this kind of data would require an unreasonable amount of work, and provide little to no value because the results for the individual projects would be project-specific and could not be compared. The Wortschatz project (Quasthoff and Richter (2005)) at Leipzig University generates a lot of annotation data and word statistics based on several sentence lists collected from online resources. The sentence lists can be considered the primary data, while everything else – including indices for the primary data – can be considered secondary data. Table 2 shows the relation between the Volumes of primary and secondary data based on the three samples *deu_mixed_2011*, *deu_news_2011* and *deu_newscrawl_2011*. The information was compiled based on information given by a server administrator with direct access to the databases.

Table 2: Primary vs secondary data Volume (*Wortschatz*).

Data Set	Primary Data (Bytes)	Secondary Data (Bytes)
deu_mixed_2011	37, 270, 576, 048	517, 020, 294, 364
deu_news_2011	3, 672, 898, 564	59, 421, 534, 187
deu_newscrawl_2011	3, 735, 178, 336	222, 879, 231, 073

The values in the table are not comparable to each other because each data set includes different sets of database tables. This is not an issue because the purpose is only to illustrate that secondary data tends to be of more Volume than primary data.

Combined with the trend for increased interoperability and research infrastructures that may store and provide annotations that would have been considered as temporary data in project-specific workflows, it may even be possible that exponential Volume growth occurs in the near future because of further annotations that are based on or caused by existing annotations.

It can be stated that secondary data itself can qualify as a Volume problem because text annotation can increase the amount of meta information that is attached to any piece of text data without limit, and therefore, the Volume can be inflated indefinitely. Estimating whether or not this would result in Big Data sized document collections would be speculation. Yet, this work proposes that it is unlikely that future document collections will include every piece of annotated information in their documents because it makes the documents harder to read, and the information may even be contradictory to each other. It is more likely and reasonable that text passage references are used to link annotation results to text passages and between external services.

1.2 Variety

Variety is about the different types and formats of data sets. Types include more broad differentiations like *audio*, *video*, or *sensory data* and also different file types for each media type like *mp3*, *wav*, and *flac* for audio files. Since the context of this work is text-oriented digital humanities, the types of data are already relatively limited but still include many file types – like tex, txt, xml, doc, csv, pdf, and many more – with specific characteristics.

Other layers of complexity in Variety are differences in markup formats for a specific file type – like different XML schemas – and a vast number of workflows and access methods for data. This indicates that the Big Data issue Variety is similar to the increasing need for interoperability that is described in Section 2 and is very relevant in the context of text-oriented digital humanities.

1.3 Velocity

Velocity describes the processing speed and is especially significant because it has a direct impact on the end-user experience while the other issues are generally only problematic for the service provider. For instance, a navigation system that calculates the best route based on sensory information about the current traffic would not be usable if this calculation requires several hours of processing time. More academic use cases are workflows that include a lot of experimental parameter permutation or the creation of domain-specific training data sets for neural networks and machine learning.

A very common way to increase the processing speed of a workflow or algorithm is to parallelize it by dividing it into subsets of problems that are independently solved by different threads or computers in a network cluster and then combining their results. Parallelization of algorithms is an issue that is far from trivial and in some cases may be counter-productive or even impossible to implement because certain workflows can not be divided into independent sub problems. Specific tasks in the text-oriented digital humanities – for example, citation analysis – can be parallelized and provide interesting research questions with regard to Velocity.

1.4 Veracity

Veracity refers to the quality and trustworthiness of data and is especially relevant in the context of sensory data where it can be a complex problem to distinguish between a correctly measured anomaly and a malfunction of a sensor. This can result in reduced efficiency and in financial losses as described in Dienst and Beseler (2016). Optical Character Recognition (OCR) can be considered as a complex Veracity-related problem in the context of text-oriented digital humanities. This observation is supported by the conclusions of Chaudhuri et al (2017). Nuances that distinguish certain letters can be hard to interpret correctly by a computer. Since OCR often has to work with documents that were not created digitally, problems like handwriting and unwanted image artefacts have to be considered. Even a comparatively high accuracy of 95% implies that every 20th character was guessed wrongly, which correlates to six mistakes in this sentence.

1.5 The Big *V*s and Digital Humanities

A problem can be more or less characterized as Big Data the more or less complex it is as regards to one or many of the Big Data *V*s. This especially implies that a problem does not necessarily have to include particularly large sets of data to be considered Big Data. The different aspects can be related to or influence each other. A relatively small data set that needs to be processed exceptionally fast is also a Big Data problem and Veracity can become decreasingly or increasingly important with increasing Volume, depending on the use case. A larger data set can decrease the impact of individual errors but also increase their absolute number in case of a systemic problem. This work argues that the following relations between the *V*s and the digital humanities can be observed:

- Volume is an issue that does exist with regard to secondary data but generally not as prominent as in other data related contexts and domains.
- Velocity and Veracity can be problematic in specific tasks in citation analysis (time effectiveness) and digital humanities like OCR (Veracity).
- Variety can be mapped to interoperability, a well known and universal issue in the digital humanities.

The following section illustrates, why interoperability or Variety is an especially complex issue in such a broad field of the digital humanities.

2 Interoperability (Variety)

Interoperability in the context of this work means the ability to interchange or reuse tools and data sets between different (research) projects. The Oxford Dictionary 2016 defines interoperability as “The ability of computer systems or software to exchange and make use of information” (Oxford Dictionary (2016)). Three technical aspects are relevant to the exchange of functions and data sets: *Tools & workflows* must understand the data, *data types & markup* must be understandable by the tools, and *data availability & access* must be provided.

2.1 Tools & workflow Variety

Many projects in the text-oriented digital humanities can be characterized as specialized solutions that are not generally applicable to other research projects as e.g. Perseus (Smith et al (2000)), *Das Deutsche Textarchiv (Geyken et al (2011))*, and The Parallel Bible Corpus (Mayer and Cysouw (2014)). They use existing or newly created technologies to provide project-specific solutions for their project-specific data sets, including the use of publicly available tools like source code repositories (Perseus) as well as hand-crafted solutions (*Das Deutsche Textarchiv*, Parallel Bible Corpus). Tool reuse can be complicated because of domain-specific circumstances. For instance, it is not unusual to use a whitespace-based word tokenizer in Latin-based languages, which cannot be applied to Chinese texts. There may also be the case that individual tasks in a workflow are considered to be solved more easily using an improvised script instead of investing the effort to evaluate already existing solutions. The result is a set of workflows that consist of an increasingly bigger set of hand-crafted project-specific programs.

The general consequence is a heterogeneity of technical solutions which makes it even harder for future researchers to find the tool combinations that are potentially useful for a given research problem. This issue is well-known in the digital humanities community as evidenced by the increasing popularity of digital infrastructures and archival projects like CLARIN (Hinrichs and Krauwer (2014)) and Das Digitale Archiv NRW (Thaller (2013)).

With the increasing familiarity, acceptance, generality, and usability of existing tools and frameworks, this variety of (potentially redundant) workflows will probably decrease over time. Source code repositories like Github are already an established technical basis for collaborative text-editing workflows[5] and mentions of natural language processing tools like the Part-of-Speech Tagger from the Stanford Natural Language Group (commonly referred to as the Stanford Tagger, Manning et al (2014)) rarely require further explanation. Yet, due to domain and context-specific requirements and also the fact that tool implementers are often motivated to try out and provide new solutions with their individual set of advantages and disadvantages, this workflow variety will probably evolve but never completely disappear, for examples, see the justifications for the toolkits that are offered by almost every Natural Language

[5] See https://github.com/PerseusDL or https://github.com/tillgrallert/digital-muqtabas.

Processing group. It is unlikely that a complicated field like the text-oriented digital humanities with its vast variety of research questions and potentially incompatible parameter configurations can be covered by a comprehensive "Jack of all trades"-kind of solution. It can also be argued that this would not be a desirable scenario since a variety of solutions can be expected to be more flexible and promote improvements by innovation. Even established tools and workflows can be expected to change over time due to updates and technical improvements or complete paradigm shifts like the currently emerging trend for workflow parallelization.

2.2 Data type & markup Variety

It can be counter-productive not to use established text-markup formats because the specification of a project-specific and competent format requires significantly more effort than the reuse of an existing one. Additionally, since formats like TEI/XML and DocBook already provide comprehensive sets of domain-specific features, it is hard to find acceptance and curiosity for new text markup formats in the research and tool development communities. It is more likely that future researchers will be trained in established markup formats and use or extend these for their purposes as, for example, described in Kalvesmaki (2015). Tool compatibility increases the value of a published data set, and therefore, it can be expected that this aspect will develop toward more interoperable data sets in established formats without further external intervention.

2.3 Data availability & access Variety

Access to data sets in the text-oriented digital humanities is generally provided through project-specific websites and solutions, including zipped data dumps (e.g. Textgrid (Neuroth et al (2011)), German Political Speeches (Barbaresi (2012))), source code repositories (e.g. Digital Muqtabas (Grallert (2016)), Perseus), and website-specific catalogues or search forms (e.g. Das Deutsche Textarchive, Parallel Bible Corpus). There does not exist a widely accepted solution for a universal interface for text data. The argument can be made that such a solution could not already be implemented because an application-

independent reference & retrieval system for text data did not exist. Text data retrieval systems like archives or website catalogues are not designed to be reusable because they are not meant to provide the basis for other systems but instead, a context-specific way to retrieve data. For example, the search catalogue that serves the data from the Parallel Bible Corpus is not designed to be also able to serve the data from *Das Deutsche Textarchiv*. Therefore, the data references can be expected to be not compatible with other projects.

Application-independent reference systems like ISBN (Griffiths (2015)) or DOI (Paskin (2010)) provide reusable identifiers for text resources but do not serve data in any way. They refer to the electronic resource as a whole, which typically correlates to one file or document while the Canonical Text Service (CTS) protocol (Smith (2009)) extends this principle to individual text passages.

This aspect has good potential for improvement. Text referencing and retrieval systems can be combined to provide access to data in an application-independent way as it is already done for complete resources as soon as a reference system like ISBN is integrated into a data archive. Adapting this principle to text passages and combining it with a retrieval web service – as it is done with the CTS implementation described in Tiepmar (2018) – can significantly increase interoperability across projects.

3 Conclusion

In summary, it can be stated that Big Data is a complex issue, especially when it is considered in a broad domain like digital humanities, even if it is restricted to the text oriented areas of this field. This paper argues that the trivial assumption that Big Data requires large data sets is not necessarily correct in this context and that other aspects and especially the issue of interoperability may be more relevant. It also shows that focusing only on volume related data aspects may result in ignorance against a significant number of potentially interesting use cases. Interoperability is further divided into three aspects and it is shown that one of them - data availability & access - shows huge potential for significant improvements. This paper lists numerous practically relevant research problems that can be considered as Big Data without requiring large data sets and in the process provides useful starting points and arguments for interested researchers that want to work in this area.

Acknowledgements Part of this work was funded by the German Federal Ministry of Education and Research within the project ScaDS Dresden/Leipzig (BMBF 01IS14014B).

References

Barbaresi A (2012) German political speeches – corpus and visualization (2nd release). In: Poster Session of the German Linguistic Society, Special Interest Group on Computational Linguistics (DGfS-CL), German Linguistic Society, Special Interest Group on Computational Linguistics (DGfS) / Open Archive of Human and Society Sciences (HAL), Frankfurt (Germany) / Paris (France), URL https://halshs.archives-ouvertes.fr/halshs-00677928

Chaudhuri A, Mandaviya K, Badelia P, Ghosh SK (2017) Optical Character Recognition Systems for Different Languages with Soft Computing. Springer International Publishing, Cham (Switzerland). DOI 10.1007/978-3-319-50252-6

Dienst S, Beseler J (2016) Automatic Anomaly Detection in Offshore Wind SCADA Data. In: Win Europe Summit Conference 2016, University of Leipzig / Global Tech I Offshore Wind GmbH, Leipzig / Hamburg (Germany), URL https://windeurope.org/summit2016/conference/submit-an-abstract/pdf/626738292593.pdf

Geyken A, Haaf S, Jurish B, Schulz M, Steinmann J, Thomas C, Wiegand F (2011) Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. In: Digitale Wissenschaft – Stand und Entwicklung digital vernetzter Forschung in Deutschland, Schomburg S, Leggewie C, Lobin H, Puschmann C (eds), Marketing des Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (hbz), Cologne (Germany), p. 157–161, URL https://hbz.opus.hbz-nrw.de/frontdoor/index/index/docId/206

Grallert T (2016) Digital Muqtabas: An open, collaborative,and scholarly digital edition of Muhammad Kurd Ali's early Arabic periodical Majallat al-Muqtabas (1906–1917/18). URL https://github.com/tillgrallert/digital-muqtabas

Griffiths S (2015) ISBN: A History. NISO's Information Standards Quarterly, Summer & Fall 2015 27(2):46–48, URL https://groups.niso.org/publications/isq/v27no2-3/Griffiths/

Hinrichs E, Krauwer S (2014) The clarin research infrastructure: Resources and tools for ehumanities scholars. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Chair) NCC, Choukri K, Declerck T, Loftsson H, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S (eds), European Language Resources Association (ELRA), Reykjavik (Iceland), p. 1525–1531, URL http://www.lrec-conf.org/proceedings/lrec2014/index.html

Kalvesmaki J (2015) Three Ways to Enhance the Interoperability of Cross-References in TEI XML. Symposium on Cultural Heritage Markup, Washington, DC (USA), vol. 16, DOI 10.4242/BalisageVol16.Kalvesmaki01

Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D (2014) The Stanford CoreNLP Natural Language Processing Toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Baltimore, MD (USA), p. 55–60, DOI 10.3115/v1/P14-5010, URL http://aclweb.org/anthology/P14-5010

Mayer T, Cysouw M (2014) Creating a massively parallel Bible corpus. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Chair) NCC, Choukri K, Declerck T, Loftsson H, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S (eds), European Language Resources Association (ELRA), Reykjavik (Iceland), p. 3158–3163, URL http://www.lrec-conf.org/proceedings/lrec2014/index.html

Neuroth H, Lohmeier F, Smith KM (2011) University of Edinburgh Library Learning Services, Edinburgh (UK). vol. 6, p. 222–231, DOI 10.2218/ijdc.v6i2.198

Oxford Dictionary (2016) Definition of interoperability in english: Interoperability. In: Oxford Dictionaries, Oxford University Press, URL https://en.oxforddictionaries.com/definition/interoperability

Paskin N (2010) Digital Object Identifier (DOI®) System, . Tertius Ltd., Oxford (UK), URL http://www.doi.org/overview/080625DOI-ELIS-Paskin.pdf

Quasthoff U, Richter M (2005) Projekt Deutscher Wortschatz. Babylonia 15(3):33–35, Babylonia / Fondazione Lingue e Culture, Bellinzona / Comano (Switzerland), URL http://babylonia.ch/de/archiv/anni-precedenti/2005/nummer-3-05/projekt-deutscher-wortschatz/

Smith DA, Rydberg-Cox JA, Crane G (2000) The Perseus Project: a digital library for the humanities. Literary and Linguistic Computing 15(1):15–25, DOI 10.1093/llc/15.1.15

Smith DN (2009) Citation in classical studies. 3(1)The Alliance of Digital Humanities Organizations (ADHO), URL http://www.digitalhumanities.org/dhq/vol/3/1/index.html

Thaller M (2013) Das Digitale Archiv NRW in der Praxis – Eine Softwarelösung zur digitalen Langzeitarchivierung. Kölner Beiträge zu einer geisteswissenschaftlichen Fachinformatik, Band 5,Verlag Dr. Kovač, Hamburg

Tiepmar J (2018) Implementation and Evaluation of the Canonical Text Services Protocol as Part of a Research Infrastructure in the Digital Humanities. PhD thesis, Leipzig University / Leipzig University Library, Leipzig, URL http://nbn-resolving.de/urn:nbn:de:bsz:15-qucosa2-212926